

01

# Web Scrapping XPATH

Club de Programación Creativa

Por Alejandro Ramos

@arhcoder



## 01 Internet

Conceptos básicos

## 02 HTML

Estructura de sitios web

## 03 XPATH

Extrayendo de un HTML

## 04 Código

Aplicando scraping





01

Internet



01

# ¿Qué es Internet ?



01

¿Qué es  
Internet ?

RED DE REDES

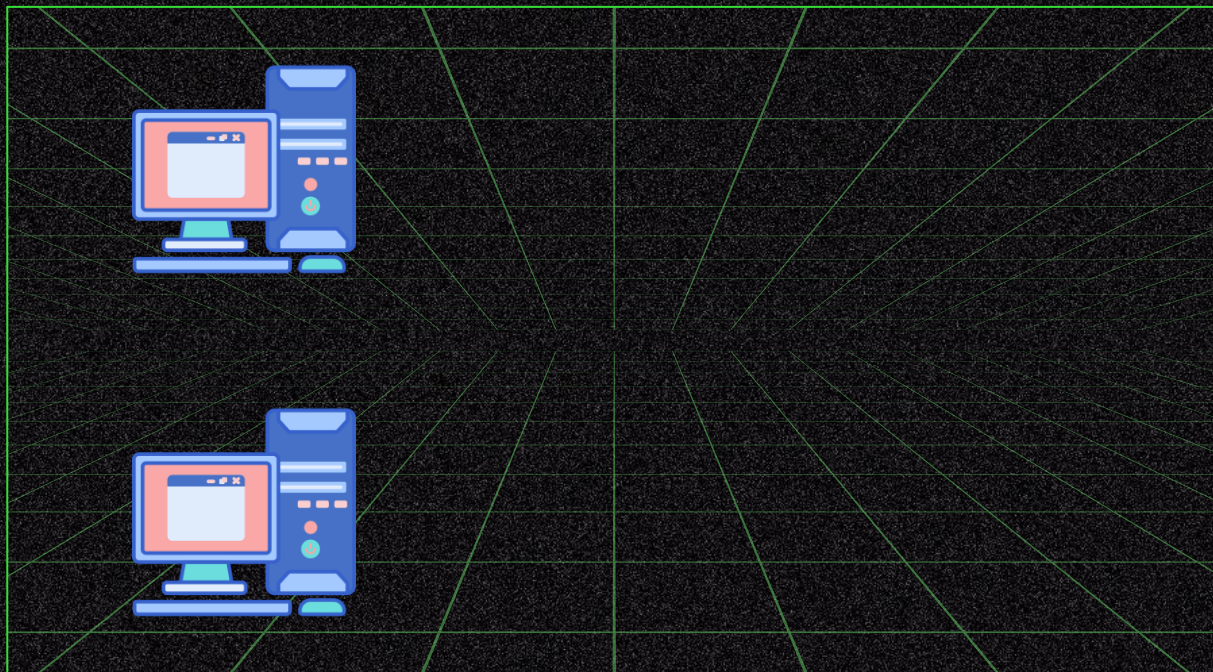


**LA**

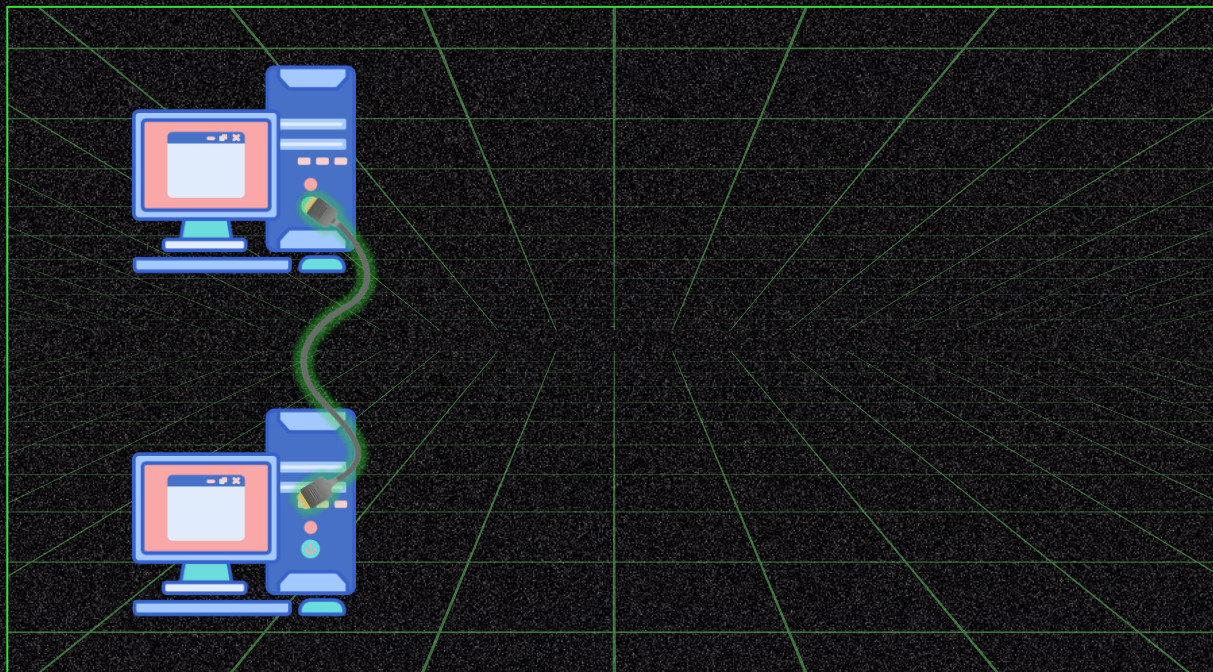
**¿Qué es  
Internet ?**

**RED DE REDES**







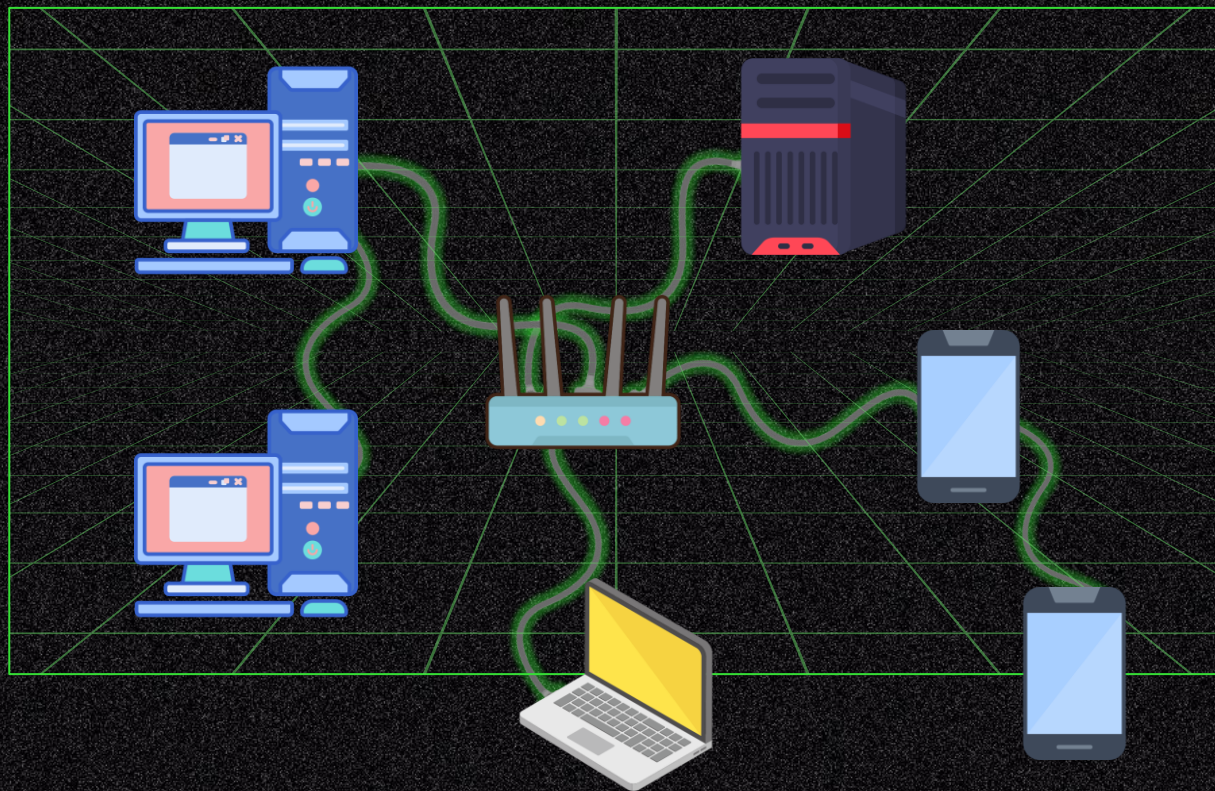






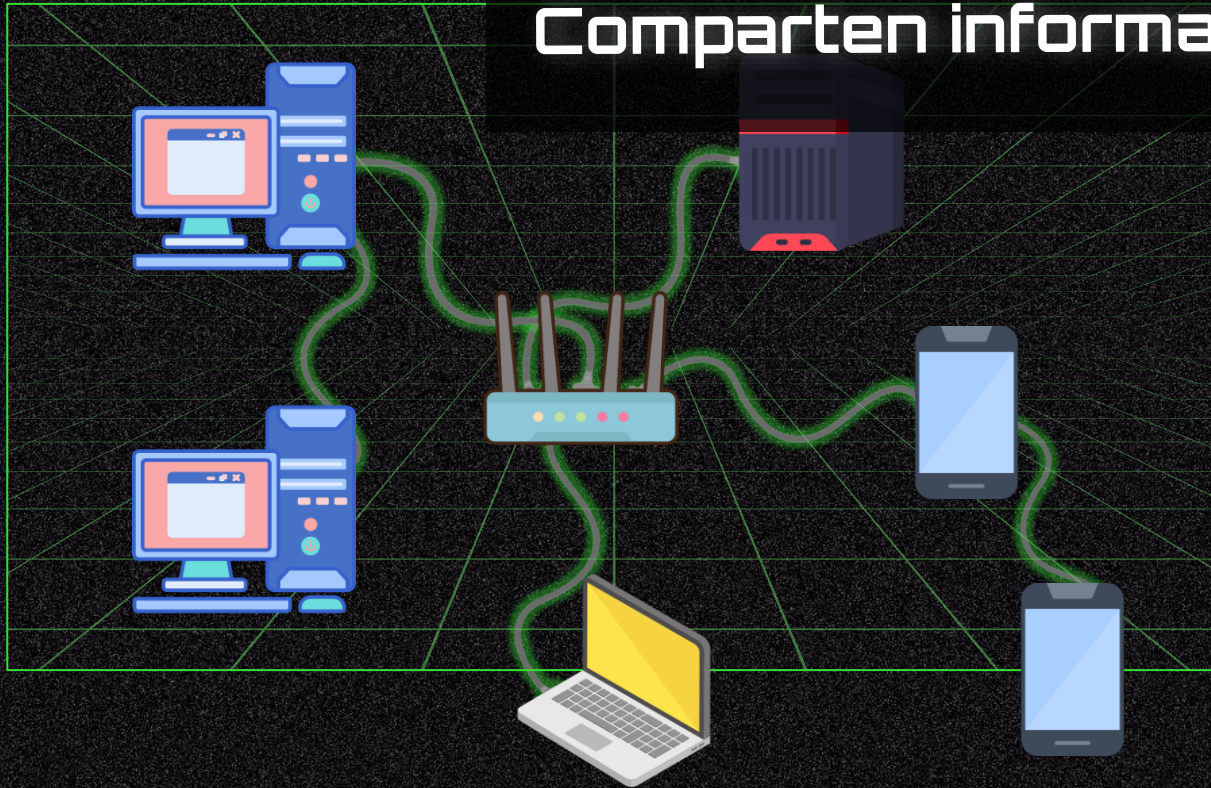
# Red de Computadoras





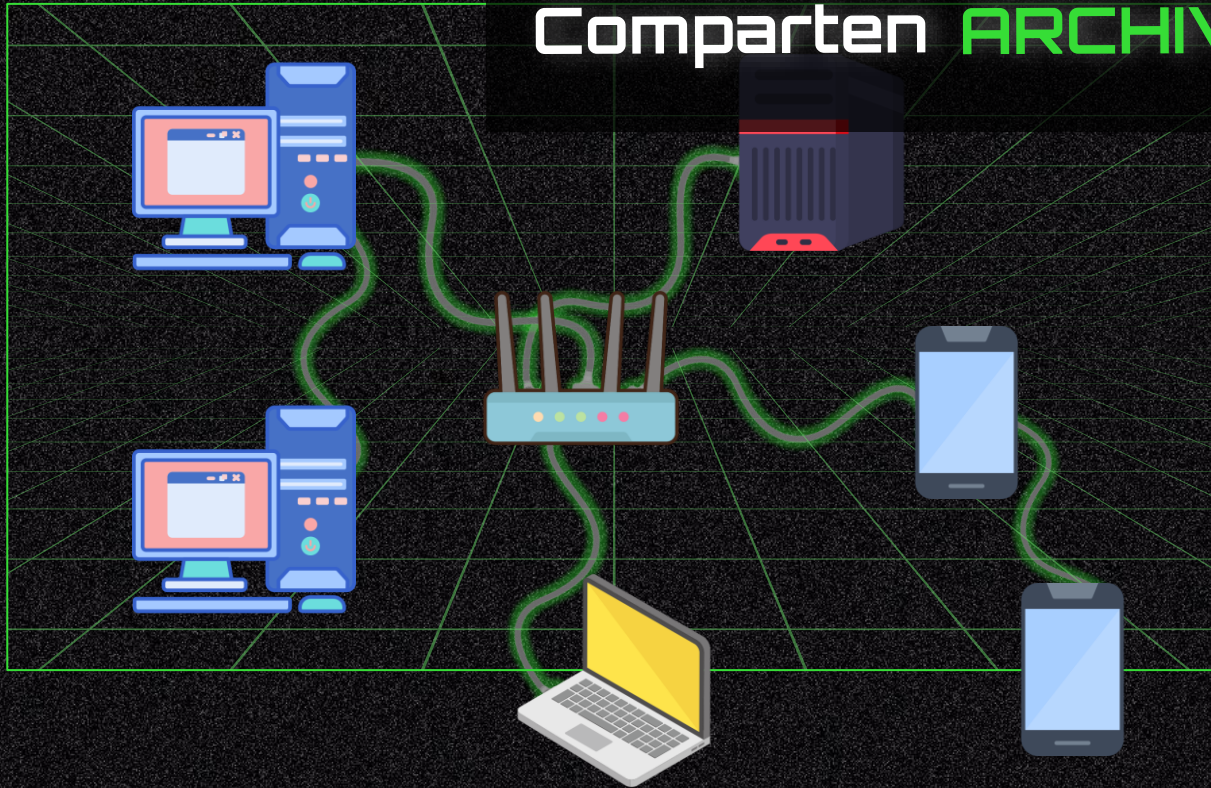


# Comparten información





Comparten **ARCHIVOS**

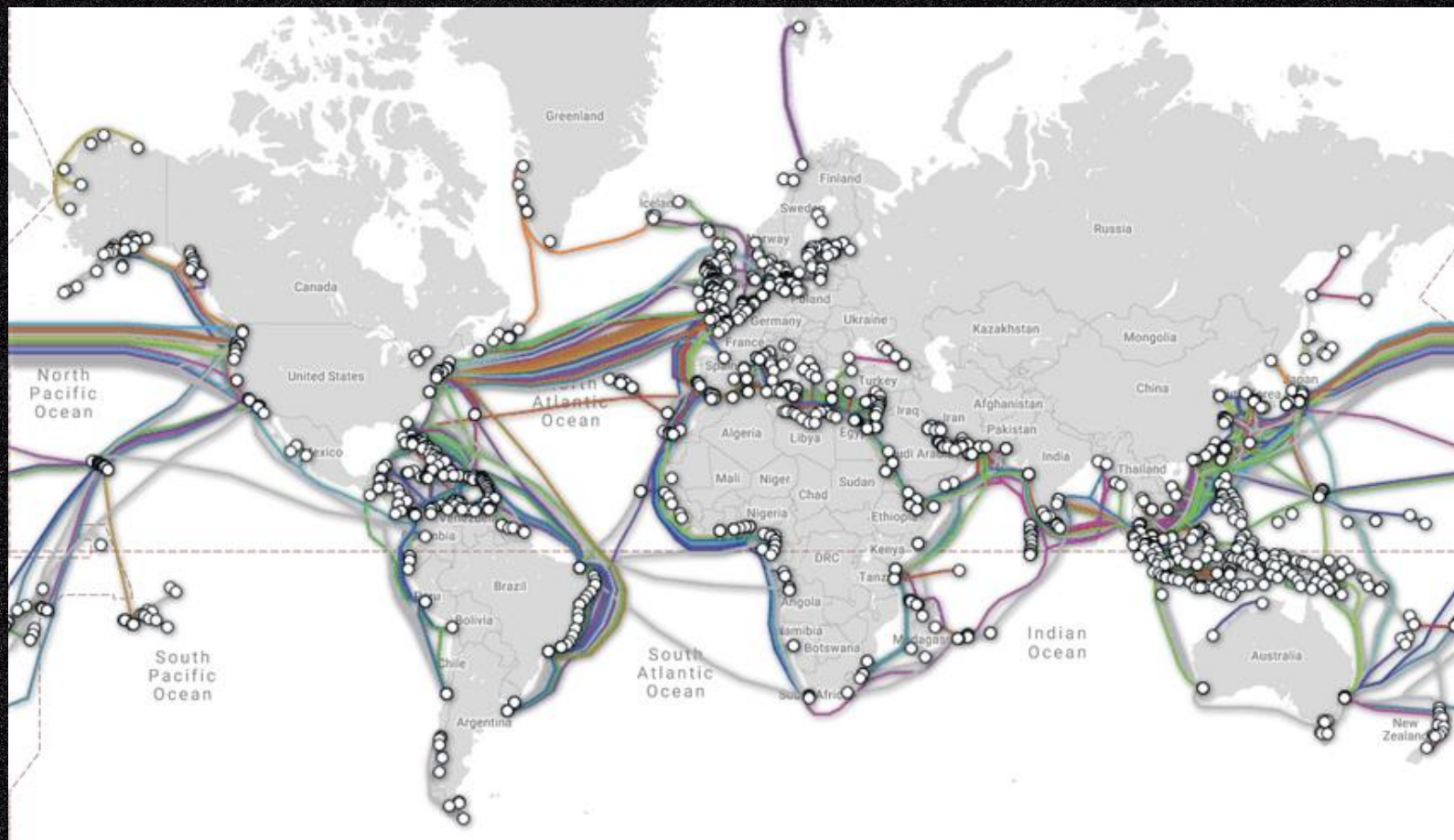




¿Qué es  
LA Internet ?

RED DE REDES





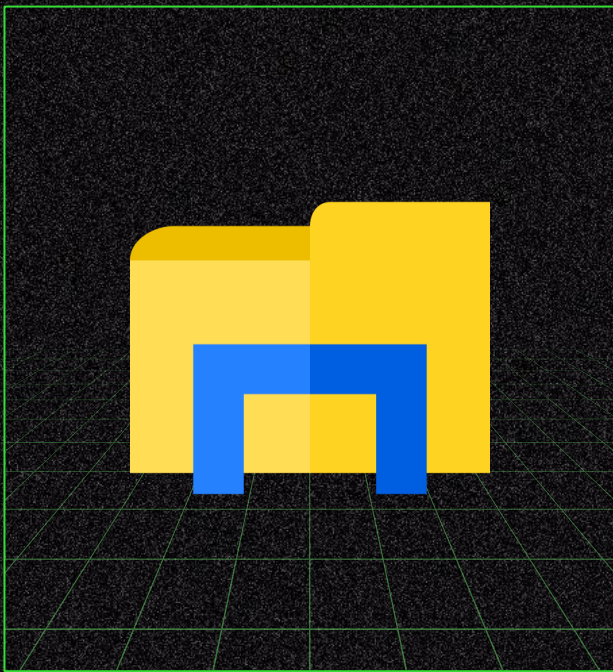


“Internet” es sólo uno; es la red más grande de computadoras en el mundo

Los dispositivos conectados pueden; **compartir** archivos o **recibir** archivos, y según su función se les llama: **servidores** o **clientes**, respectivamente







≈







Navega entre archivos  
de **MI COMPUTADORA**



Navega entre archivos de  
**OTRAS COMPUTADORAS**





Navega entre archivos  
de **MI COMPUTADORA**

Navega entre archivos de  
**OTRAS COMPUTADORAS**



Navega entre archivos de  
**OTRAS COMPUTADORAS**


Navega entre archivos  
de **MI COMPUTADORA**



Aula Virtual Licenciatura: Log in

aulavirtual.uaa.mx/login/index.php

FacebookWhatsAppYouTubeARHLinkedInHotmailEnsecodeInstagramGitHubDriveFlatconTraductorPDFAPA

UNIVERSIDAD AUTÓNOMA DE AGUASCALIENTES


**aula VIRTUAL**  
Educación Superior

Log in using your account on:

Username

@edu.uaa.mx

Password

Entrar

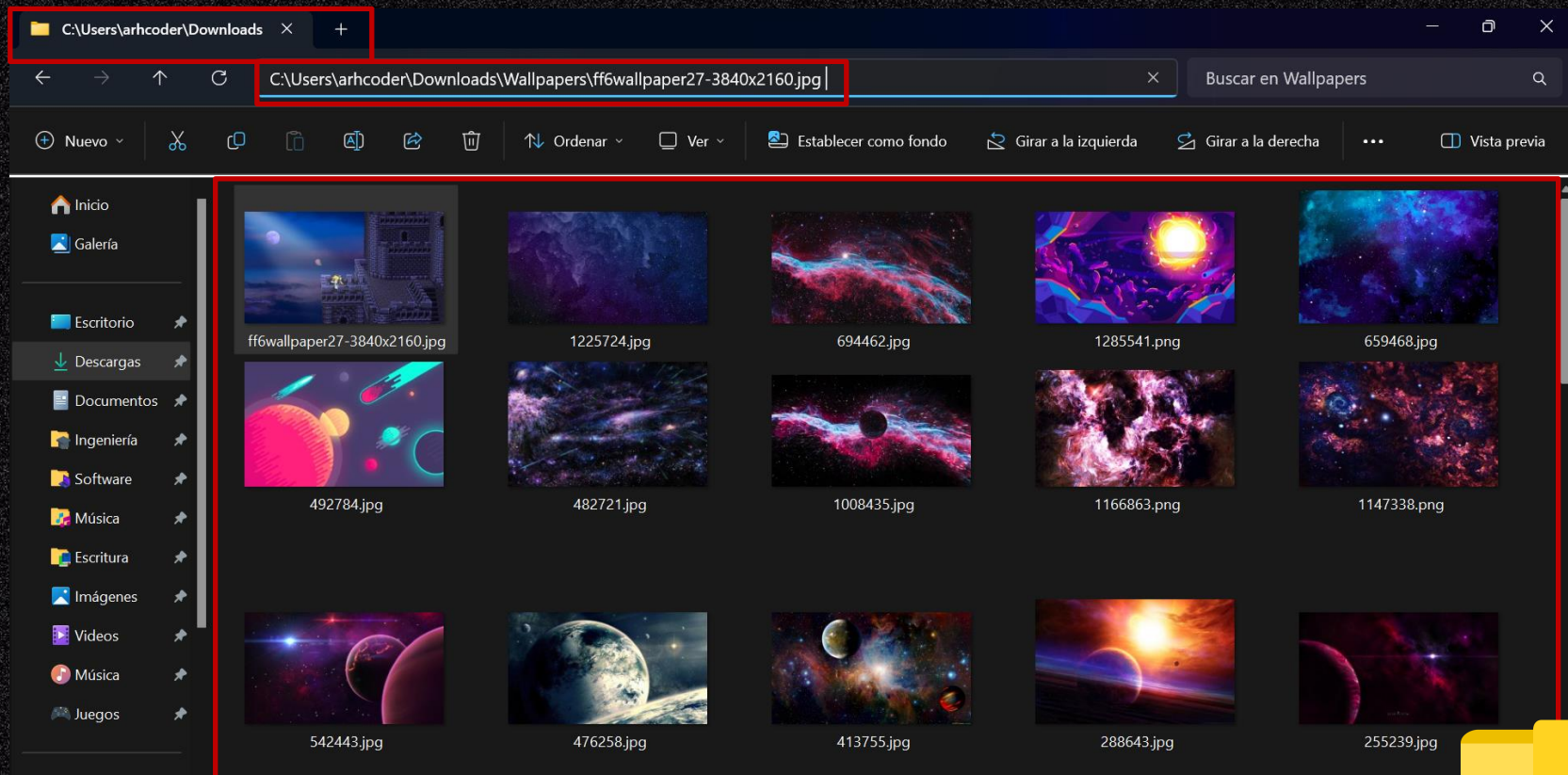
Cookies must be enabled in your browser

Atento aviso a profesores

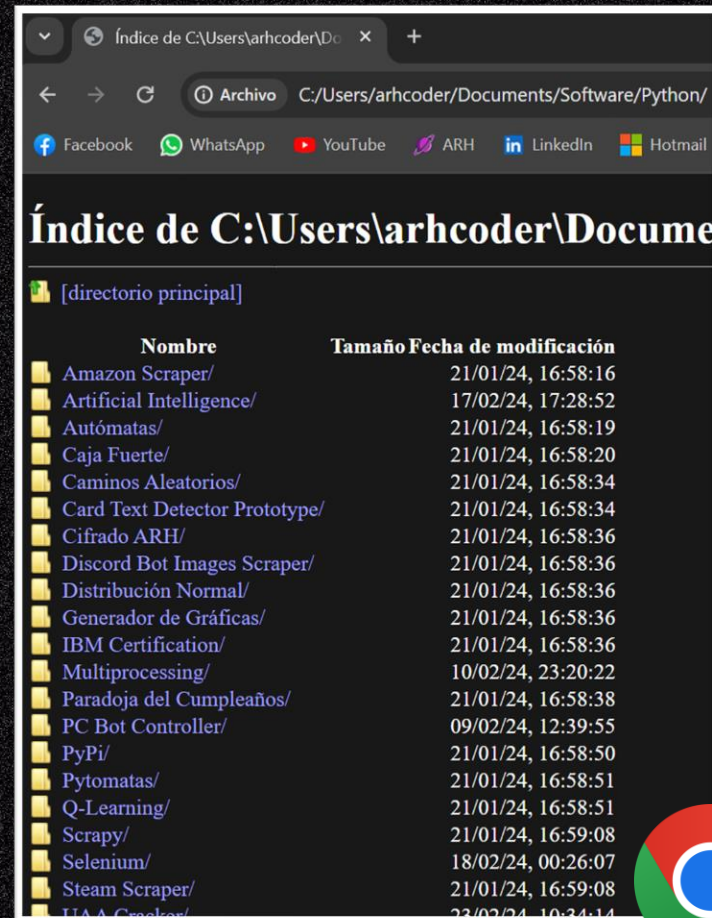
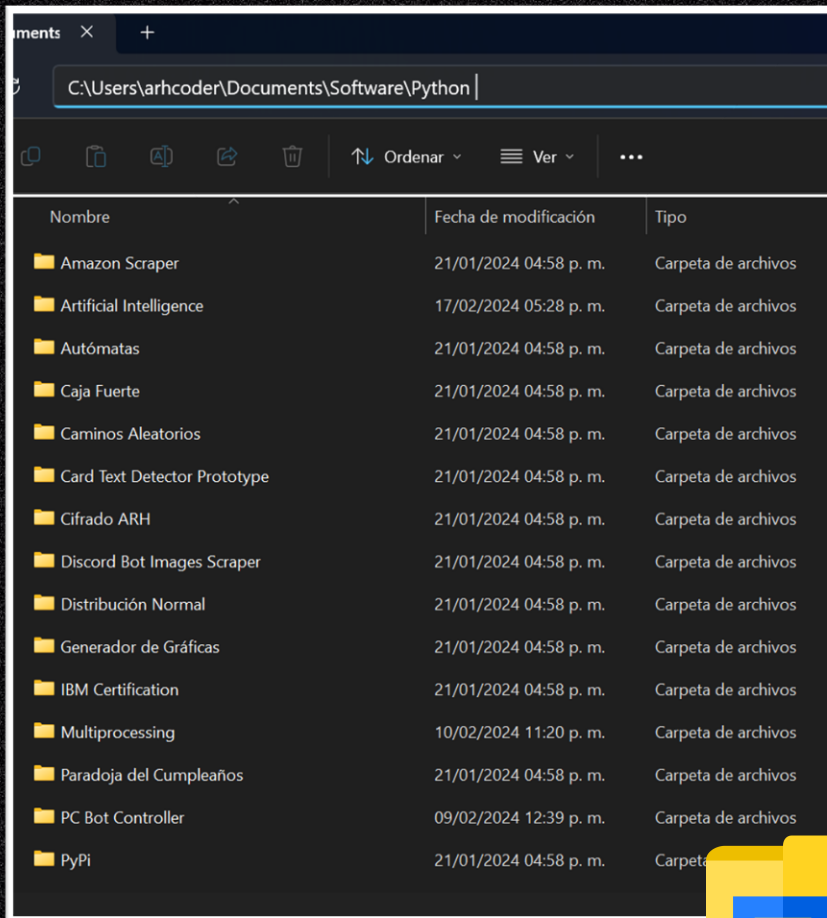
- "Videotutoriales. Uso y Manejo de Aula Virtual" da clic aquí
- Para acceder a Cursos 2021, 2022 y 2023 da clic aquí
- Si requieres restaurar un curso histórico en la nueva plataforma da













# Al intentar entrar en un sitio web

1. **MI computadora** se conecta => módem  
=> servicio de internet => servicios de internet del mundo
2. El **url** indica la dirección de **LA computadora** en el mundo
3. Se busca la ruta a **LA computadora**
4. El navegador solicita a **LA computadora** información
5. **LA computadora** envía una respuesta





# Al intentar entrar en un sitio web

1. **MI computadora** se conecta => módem  
=> servicio de internet => servicios de  
internet del mundo
2. El **url** indica la dirección de  
**computadora** en el mundo
3. Se busca la ruta a **LA computadora**
4. El navegador solicita a **LA  
computadora** información
5. **LA computadora** envía una respuesta

# request





# Al intentar entrar en un sitio web

1. **MI computadora** se conecta => módem  
=> servicio de internet => servicios de  
internet del mundo

2. El **url** indica la dirección de **LA computadora** en el mundo

3. Se busca la ruta a **LA computadora**

4. El navegador solicita a **LA computadora** información

5. **LA computadora** envía una respuesta

# HTTP request





# Al intentar entrar en un sitio web

1. **MI computadora** se conecta => módem  
=> servicio de internet => servicios de  
internet del mundo

2. El **url** indica la dirección de **LA computadora** en el mundo

3. Se busca la ruta a **LA computadora**

4. El navegador solicita a **LA computadora** información

5. **LA computadora** envía una respuesta

## HTTP request

**POST -**

**GET -**

**PUT -**

**DELETE -**



**Una vez que el navegador  
recibe una respuesta, la  
abre como archivo\***





02

HTML



# HTML

133 idiomas

Artículo Discusión

Leer Editar Ver historial Herramientas

## Contenidos

ocultar

### Inicio

Primeras especificaciones de HTML

### Marcador HTML

Elementos

Atributos

Etiquetas HTML básicas

Nociones básicas de HTML

Editores de textos

Aprender HTML analizando páginas reales

Historia del estándar

Accesibilidad web

Entidades HTML

**HTML**, acrónimo en inglés de *HyperText Markup Language* («lenguaje de marcado de hipertexto»), hace referencia al [lenguaje de marcado](#) utilizado en la creación de [páginas web](#). Este estándar que sirve de referencia del software que interactúa con la elaboración de páginas web en sus diferentes versiones. Define una estructura básica y un código (denominado código HTML) para la presentación de contenido de una página web, que incluye texto, imágenes, videos, juegos, entre otros elementos. Este estándar es gestionado por el [World Wide Web Consortium](#) (W3C) o Consorcio WWW, una organización dedicada a la estandarización de la mayoría de las tecnologías asociadas a la web, especialmente en lo relacionado con su escritura e interpretación. HTML se considera el lenguaje web más importante y su invención crucial para el surgimiento, desarrollo y expansión de la [World Wide Web](#) (WWW). Es el estándar que prevalece en la visualización de páginas web y es adoptado por todos los navegadores actuales.<sup>1</sup>

El lenguaje HTML se fundamenta en la diferenciación como filosofía de desarrollo. Para añadir elementos externos a una página como imágenes, vídeos o *scripts*, no se incrustan directamente en el código de la página. En su lugar, se realiza una referencia a la ubicación de cada elemento mediante texto. De este modo, la página web contiene solamente texto.

## HTML

<HTML>

```
<html>
<title>HTML</title>
<body>
This is HTML!
</body>
</html>
```

HTML



**HTML** es un lenguaje de programación creado como estándar para construir **estructuras de sitios web**

HTML se escribe en archivos y estos archivos representan **"documentos"**, tal cual como un PDF, XML, etc, son la estructura de información fija

```
18 <title>Tic, Tac, Toe!</title>
19 </head>
20
21 <body onload="reset()">
22   <div class="top-panel">
23     <div class="titular"><h1>¡TIC, TAC, TOE!</h1><
24     <div id="player-turn"><h2 id="turn-text">Inici
25   </div>
26   <div class="content">
27     <div class="tictactoe">
28       <div class="box-empty" id="box1" onclick="
29       <div class="box-empty" id="box2" onclick="
30       <div class="box-empty" id="box3" onclick="
31       <div class="box-empty" id="box4" onclick="
32       <div class="box-empty" id="box5" onclick="
33       <div class="box-empty" id="box6" onclick="
34       <div class="box-empty" id="box7" onclick="
35       <div class="box-empty" id="box8" onclick="
36       <div class="box-empty" id="box9" onclick="
37     </div>
38   </div>
39   <div class="bottom-panel">
40     <div class="scores" id="scores">
41       <div id="score-x-box">
42         <h2 id="x-score">0</h2>
45         <img src="svg/o.svg" class="score-icor
```



**HTML** se conforma de **ETIQUETAS (TAGS)**, las cuales se refieren a "elementos" que serán mostrados en pantalla; textos, imágenes, botones, etc.

Las etiquetas se colocan **dentro** de otras etiquetas y estas a su vez **dentro de otras**, generando la **estructura** de un sitio web

```
18 <title>Tic, Tac, Toe!</title>
19 </head>
20
21 <body onload="reset()">
22   <div class="top-panel">
23     <div class="titular"><h1>¡TIC, TAC, TOE!</h1><
24     <div id="player-turn"><h2 id="turn-text">Inici
25   </div>
26   <div class="content">
27     <div class="tictactoe">
28       <div class="box-empty" id="box1" onclick="
29       <div class="box-empty" id="box2" onclick="
30       <div class="box-empty" id="box3" onclick="
31       <div class="box-empty" id="box4" onclick="
32       <div class="box-empty" id="box5" onclick="
33       <div class="box-empty" id="box6" onclick="
34       <div class="box-empty" id="box7" onclick="
35       <div class="box-empty" id="box8" onclick="
36       <div class="box-empty" id="box9" onclick="
37     </div>
38   </div>
39   <div class="bottom-panel">
40     <div class="scores" id="scores">
41       <div id="score-x-box">
42         <h2 id="x-score">0</h2>
45         <img src="svg/o.svg" class="score-icor
```



```

<body>
  <header>
    <h1>Mi Página WEB</h1>
  </header>

  <section>
    <article>
      <h2>Sección 1</h2>
      <p>Este es un párrafo dentro de un artículo. También hay una imagen:</p>
      
      <button>Botoncito</button>
    </article>
  </section>

  <section>
    <article>
      <h2>Sección 2</h2>
      <p>Otro párrafo dentro de otro artículo. Aquí hay un enlace y un formulario:</p>
      <a href="https://www.google.com">Enlace a google.com</a>
      <form>
        <label for="nombre">Nombre:</label>
        <input type="text" id="nombre" name="nombre">
        <button type="submit">Enviar</button>
      </form>
    </article>
  </section>

  <footer>
    <p>&copy; 2024 Mi Página Web</p>
  </footer>
</body>

```

# Mi Página WEB

## Sección 1

Este es un párrafo dentro de un artículo. También hay una imagen:



Botoncito

## Sección 2

Otro párrafo dentro de otro artículo. Aquí hay un enlace y un formulario:

[Enlace a google.com](https://www.google.com)

Nombre:

Enviar

© 2024 Mi Página Web



# Partes de una etiqueta [TAG]

## Nombre

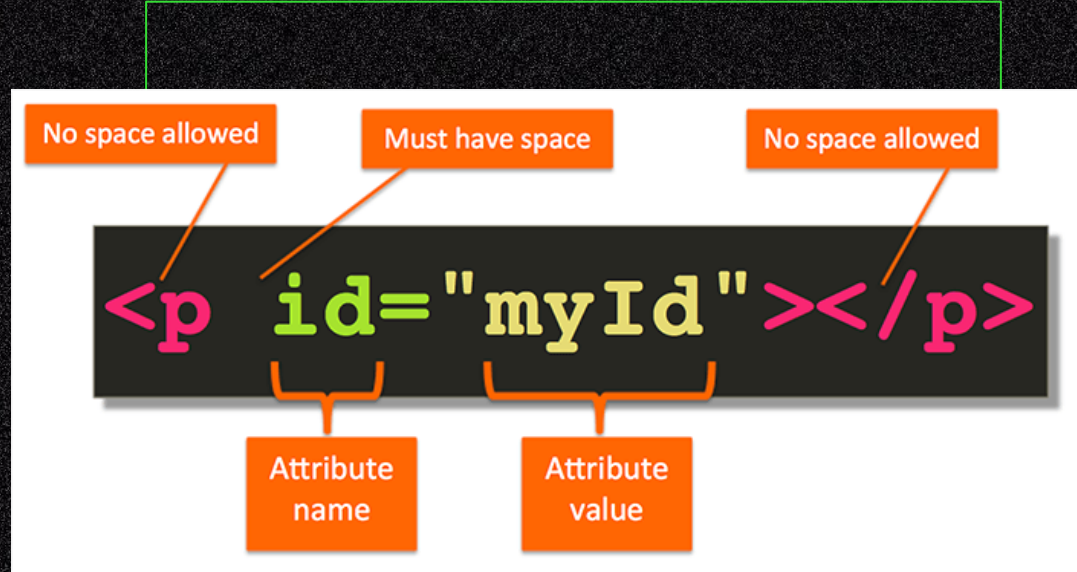
Qué tipo de objeto se ilustrará en pantalla; **img**, **p**, **button**

## Atributos

**Características** que posee la etiqueta

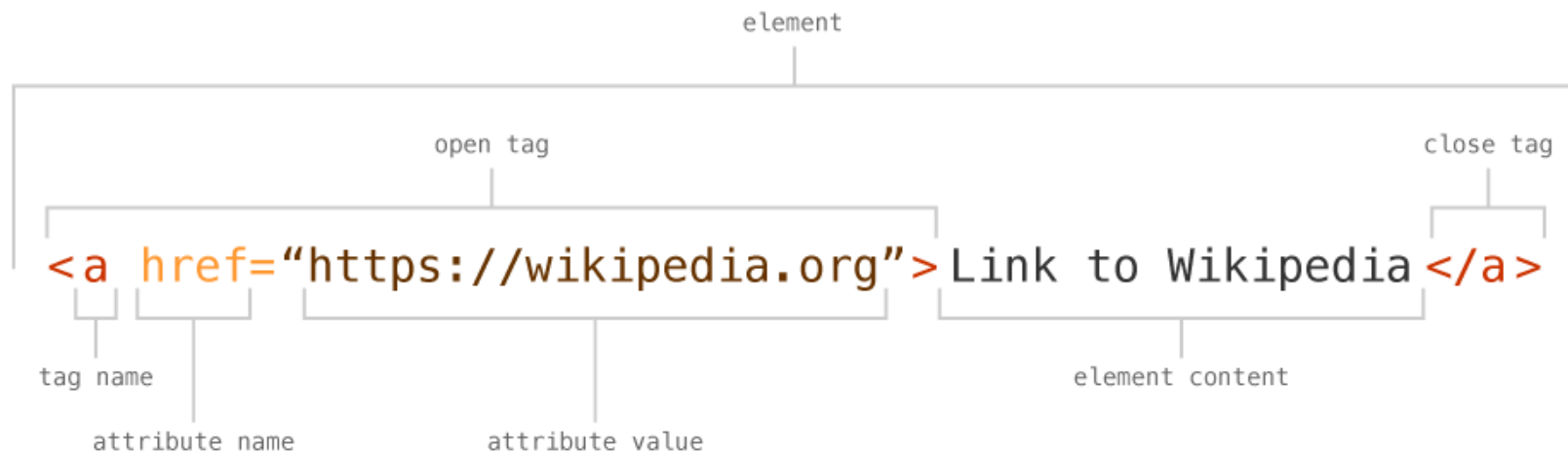
## Contenido

Lo que sea que esté entre el **<inicio>** y el **<\cierre>** de la etiqueta





# EJEMPLO DE ETIQUETA HTML





## HTML

## Contenidos

Artículo

## Discusión

Leer

Editor

Ver historial

Herramientas ▾

## Inicio

## Primeras especificaciones de HTML

## ▼ Marcador HTML

## Elementos

### Atributos

## Etiquetas HTML básicas

## Nociones básicas de HTML

## Editores de textos

## Aprender HTML analizando páginas reales

## Historia del estándar

## Accesibilidad web

## Entidades HTML

**HTML**, acrónimo en inglés de **HyperText Markup Language** («lenguaje de marcado de hipertexto»), hace referencia al [lenguaje de marcado](#) utilizado en la creación de [páginas web](#). Este estándar que sirve de referencia del software que interactúa con la elaboración de páginas web en sus diferentes versiones. Define una estructura básica y un código (denominado código HTML) para la presentación de contenido de una página web, que incluye texto, imágenes, videos, juegos, entre otros elementos. Este estándar es gestionado por el [World Wide Web Consortium](#) (W3C) o Consorcio WWW, una organización dedicada a la estandarización de la mayoría de las tecnologías asociadas a la web, especialmente en lo relacionado con su escritura e interpretación. HTML se considera el lenguaje web más importante y su invención crucial para el surgimiento, desarrollo y expansión de la [World Wide Web](#) (WWW). Es el estándar que prevalece en la visualización de páginas web y es adoptado por todos los navegadores actuales.<sup>1</sup>

El lenguaje HTML se fundamenta en la diferenciación como filosofía de desarrollo. Para añadir elementos externos a una página como imágenes, vídeos o *scripts*, no se incrustan directamente en el código de la página. En su lugar, se realiza una referencia a la ubicación de cada elemento mediante texto. De este modo, la página web contiene solamente texto.

## HTML

# <HTML>

```
<html>
<title>HTML</title>
<body>
This is HTML!
</body>
</html>
```

## UTMI





```

<ul class="vector-toc-contents" id="mw-panel-toc-list">
  <li id="toc-mw-content-text"
    class="vector-toc-list-item vector-toc-level-1">
    <a href="#" class="vector-toc-link">
      <div class="vector-toc-text">Inicio</div>
    </a>
  </li>
  <li id="toc-Primeras_especificaciones_de_HTML"
    class="vector-toc-list-item vector-toc-level-1 vector-toc-list-item-expanded">
    <a class="vector-toc-link" href="#Primeras_especificaciones_de_HTML">
      <div class="vector-toc-text">
        <span class="vector-toc-numb">1</span>Primeras especificaciones de HTML</div>
      </a>

    <ul id="toc-Primeras_especificaciones_de_HTML-sublist" class="vector-toc-list">
    </ul>
  </li>
  <li id="toc-Marcador_HTML"
    class="vector-toc-list-item vector-toc-level-1 vector-toc-list-item-expanded">
    <a class="vector-toc-link" href="#Marcador_HTML">
      <div class="vector-toc-text">
        <span class="vector-toc-numb">2</span>Marcador HTML</div>
      </a>

      <button aria-controls="toc-Marcador_HTML-sublist" class="cdx-button cdx-button--i
        <span class="vector-icon vector-icon--x-small mw-ui-icon-wikimedia-expand"><
        <span>Alternar subsección Marcador HTML</span>
      </button>

    <ul id="toc-Marcador_HTML-sublist" class="vector-toc-list">
      <li id="toc-Elementos"
        class="vector-toc-list-item vector-toc-level-2">
        <a class="vector-toc-link" href="#Elementos">
          <div class="vector-toc-text">
            <span class="vector-toc-numb">2.1</span>Elementos</div>
          </a>

```



En **WEB SCRAPING** suele  
llamársele **NODOS** a las  
etiquetas (elementos); es  
decir a los elementos que  
forman un sitio web



El proceso de **SCRAPING**  
consistirá en **obtener el**  
**HTML** de un sitio web y  
**extraer los NODOS** que  
nos interesen



# Club de Programación Creativa En YouTube



Alejandro Ramos  
@arhcoder